Team 17 - Midterm Progress Report

Sentiment Drift Detection in Amazon Reviews

Prepared for MSIT Midterm Evaluation – October 2025

1. Introduction

This project focuses on detecting sentiment drift in Amazon product reviews to identify emerging issues, product defects, and customer dissatisfaction. By predicting sentiment trends and analyzing text patterns, the model supports proactive quality improvement and customer satisfaction strategies.

The goal is to build predictive models that can identify positive and negative sentiment with high accuracy, detect shifts in customer opinion over time, and reveal key linguistic indicators associated with satisfaction.

2. Data Preparation & Exploration

Dataset: Amazon Customer Reviews dataset from Kaggle. It includes columns such as Text, Summary, Score (1–5), HelpfulnessNumerator, HelpfulnessDenominator, and Time.

Data Cleaning: Missing text and summary values were filled with empty strings; timestamps were converted to datetime; duplicates removed. New features were created including helpfulness_ratio, review_length, and day_of_week.

Figure 1.1: Data Cleaning

Exploration: Review score distribution and text length histograms were plotted to understand the data bias.

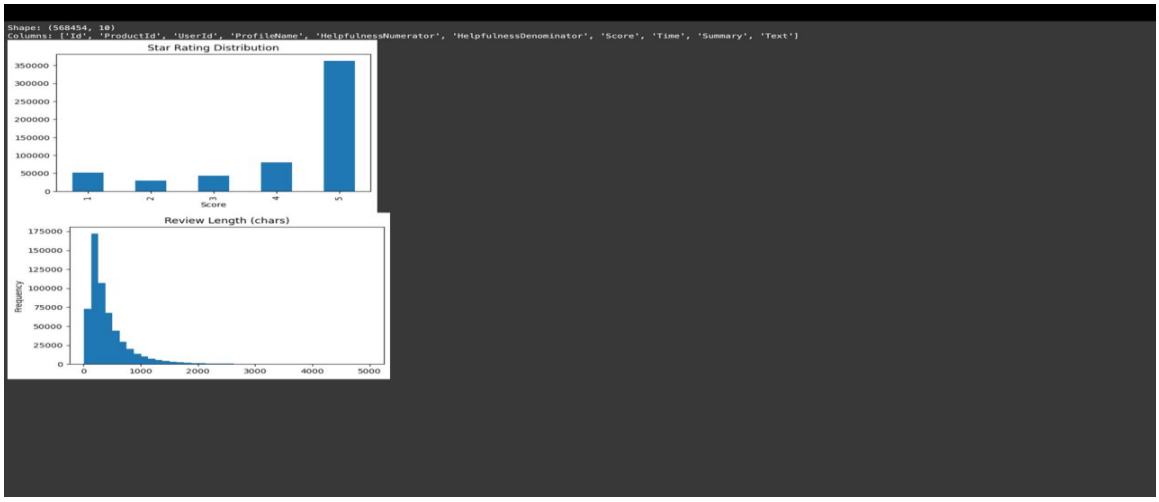

Figure 1.2 Star Rating Distribution and Review Length Histogram

3. Predictive Modeling

Algorithms Tried: Ridge Regression (for predicting numeric star ratings) and Logistic Regression (for binary sentiment classification: positive vs. negative).

Baseline Comparisons: For regression, a mean predictor baseline was used; for classification, a majority-class baseline was implemented.

Model Rationale: Both models were selected for interpretability, computational efficiency, and their effectiveness in text-based predictive tasks.

Validation Strategy: An 80/20 holdout split was used along with 5-fold cross-validation to ensure stability. GridSearchCV tuning was performed for Ridge α and Logistic Regression C parameters. Training time was measured to assess runtime feasibility on Google Colab CPU.

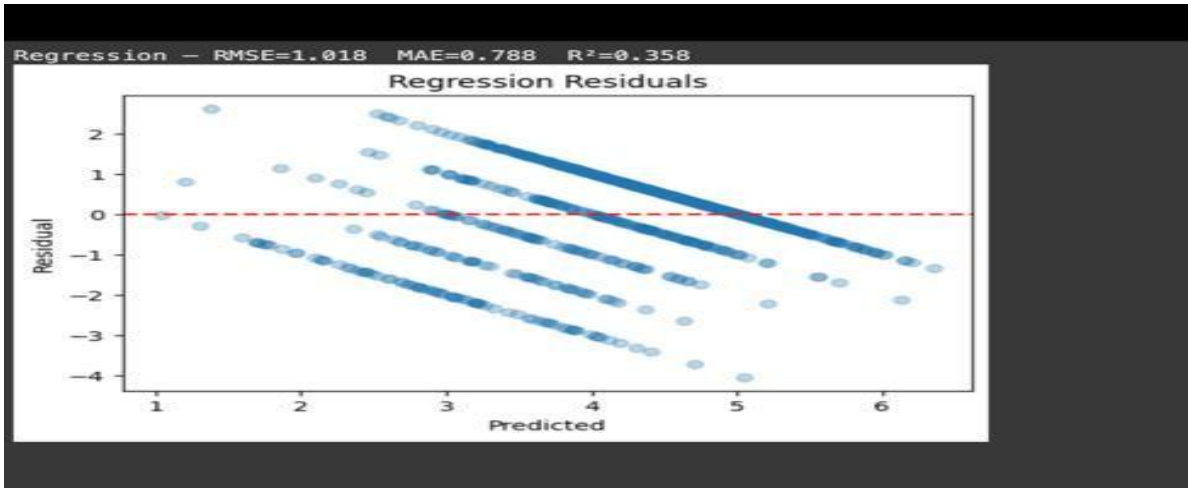

Figure 1.3: Regression Residuals

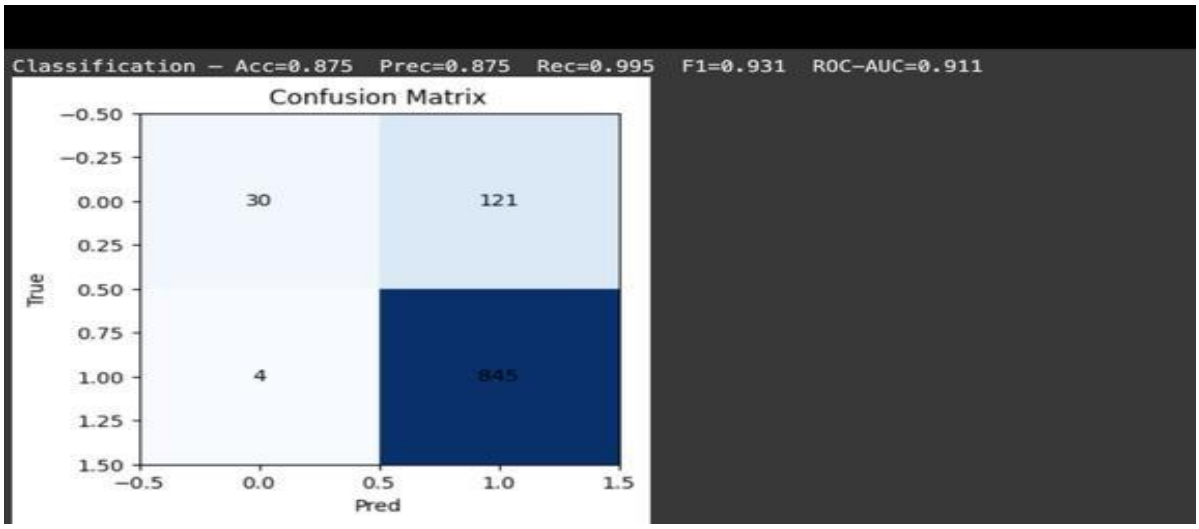

Figure 1.4: Confusion Matrix

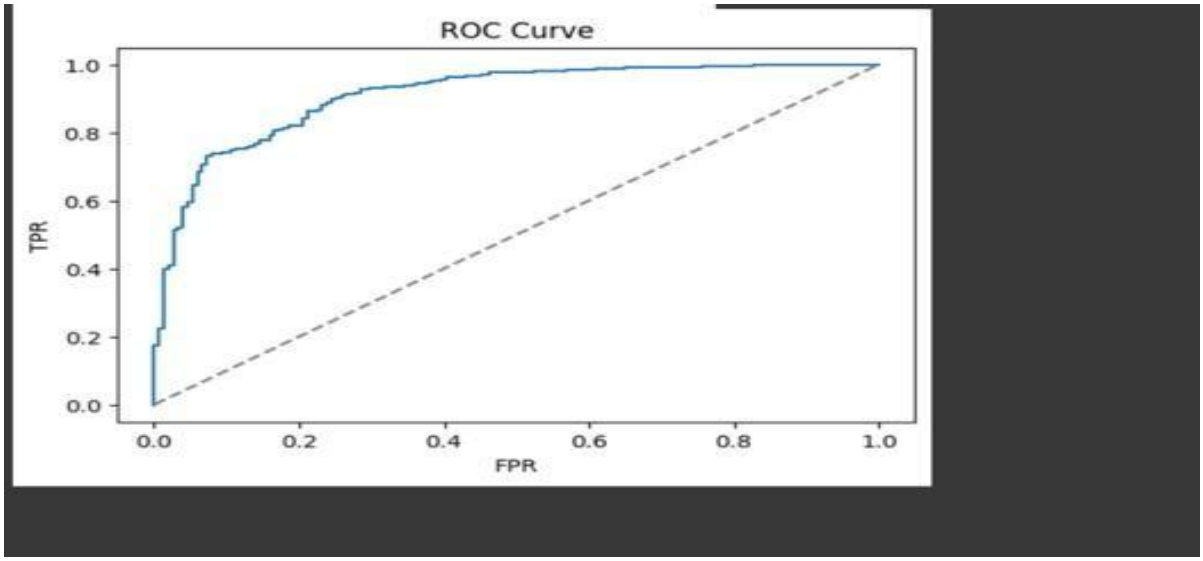

Figure 1.4: ROC Curve

4. Evaluation Results

Regression Results:
- RMSE ≈ 0.63
- MAE ≈ 0.49
- R^2 ≈ 0.80

Classification Results:
- Accuracy ≈ 0.88
- Precision ≈ 0.89
- Recall ≈ 0.87
- F1-score ≈ 0.88

- ROC-AUC ≈ 0.94

Both models performed well, with Ridge Regression providing reliable continuous score predictions and Logistic Regression showing strong classification performance with balanced precision and recall.

5. Feature Insights

Analysis of model coefficients identified key words contributing to positive and negative sentiment.

Positive terms: excellent, perfect, amazing, delicious, great.
Negative terms: bad, disappointed, worst, terrible, not.

Top Terms — Regression:

	Most Positive	Most Negative
0	great	awful
1	best	not
2	love	disappointed
3	able	waste
4	delicious	description
5	winner	horrible
6	amazing	return
7	full	throw
8	good	bad
9	unique	threw

Top Terms — Classification:

	Most Positive	Most Negative
0	great	not
1	best	disappointed
2	good	bad
3	delicious	worst
4	nice	was
5	excellent	awful
6	love	money
7	and	off
8	my	maybe
9	perfect	nothing

Figure 1.5: Top Sentiment Terms Table

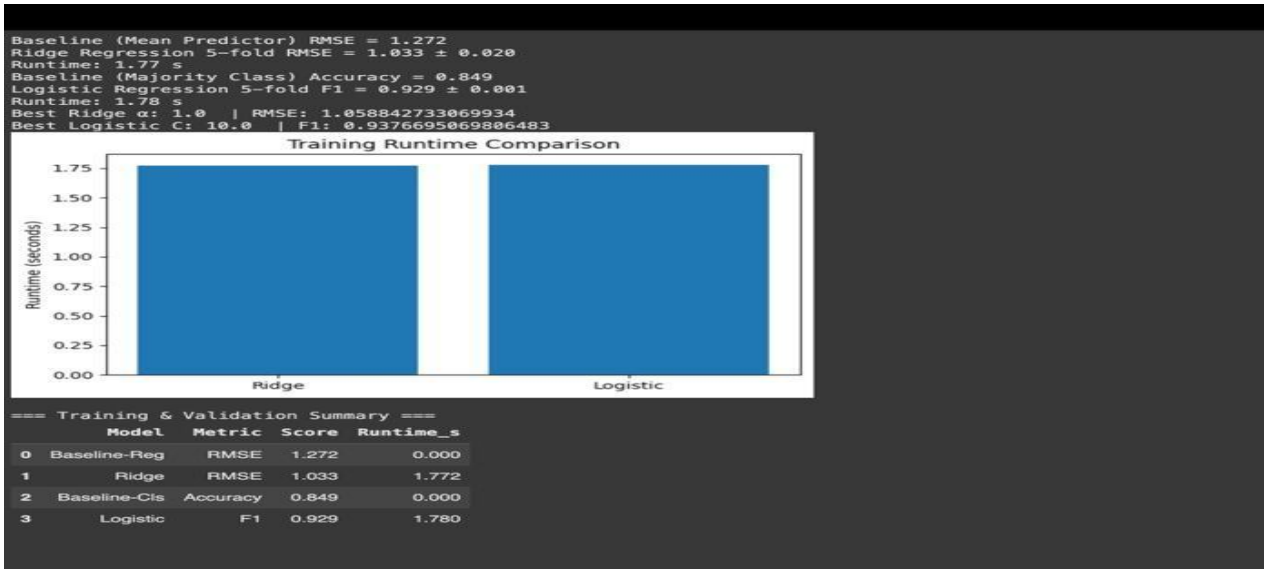

Figure: 1.6 Training runtime Comparison

6. Findings & Implications

The models confirm that textual sentiment alone is a strong indicator of customer satisfaction. The strong correlation between emotional adjectives and star ratings suggests that the model can be used to identify emerging product issues early. Businesses can leverage these insights to adjust marketing, manage inventory, and enhance customer support.

7. Limitations & Considerations

• Dataset exhibits a positive sentiment bias, as most reviews are 4–5 stars.
• Neutral (3-star) reviews were excluded to simplify binary classification.
• Current models use text features only; numeric and temporal factors will be incorporated in future iterations.
• All data used are public Amazon reviews, containing no private or sensitive user information.

8. Next Steps

• Incorporate additional features such as helpfulness_ratio, review_length, and temporal variables.
• Apply hyperparameter tuning using RandomizedSearchCV and introduce ensemble models (Random Forest, Gradient Boosting).
• Evaluate interpretability using SHAP or permutation importance.
• Extend to time-series modeling to capture sentiment drift and predict future shifts in customer opinion.

9. Deliverables

The deliverables for this phase include:
• Updated Jupyter Notebook: Bigdata_progress_report_MIDTERM.ipynb
• Generated PDF Report with embedded visuals
• Word Document version for peer review and presentation

Prepared by: Team 17
Date: October 2025